

Incentivizing Versatile Video Reasoning in MLLMs via Data-Efficient Reinforcement Learning (Supplementary Material)

A. Additional Analysis

Results on Latest Models. Tab. 1 shows that multi-task cold start brings consistent benefits to the latest Qwen3-VL-8B [1] model. We expect that multi-task RL will yield even more significant improvements, which we plan to incorporate in future versions.

Table 1. Experiments on the latest Qwen3-VL model.

Model	Video-MME			overall
	short	medium	long	
Qwen3-VL-8B-Ins	78.4	64.7	55.3	66.1
+ Ours Cold Start	78.7	64.7	56.5	66.6

Training data comparison. Cold start plays a crucial role as the initial stage of reinforcement learning. In cold start, Video-R1 trains for 1 epoch using 165k CoT data, while our method only constructs and uses 3k data for 3 epoch training. In RL, Video-R1 constructs 260k data and trains for 1200 steps, as shown in the official repository, thus using only 9.6k data. However, how to select 9.6k data from 260k data is unknown. While we directly used 5k data from NEX-T-GQA without any additional processing, as shown in Tab. 2.

Table 2. Training data comparison.

Model	Code Start	RL
Video-R1	165k	9.6k
Ours	3k	5k

B. Implement Details

Codebase. Our codebase is build on MS-Swift¹. For the multi-task cold start, we follow the standard supervised fine-tuning scripts in MS-Swift, and the detailed command is shown below:

¹<https://github.com/modelscope/ms-swift>

```

CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 torchrun
--nproc_per_node="8" \
src/open_rl/sft_three_mixture.py \
--output_dir
    ckpt/${SWANLAB_PROJECT}/${SWANLAB_NAME} \
--model_name_or_path Qwen/Qwen2-VL-7B \
--dataset_name
    data/sft_data/mix_sft_data_new_3154.json \
--deepspeed local_scripts/zero3_offload.json \
--per_device_train_batch_size 1 \
--gradient_accumulation_steps 1 \
--learning_rate 2e-6 \
--logging_steps 1 \
--bf16 \
--report_to swanlab \
--gradient_checkpointing true \
--attn_implementation flash_attention_2 \
--num_train_epochs 3 \
--run_name ${SWANLAB_NAME} \
--save_steps 100 \
--max_grad_norm 5 \
--save_only_model true \
--save_total_limit 10 \

```

For the multi-task RL training, we do not use VLLM for trajectory sampling, because using Transformers directly for sampling can ensure that the generated tokens and logit calculations are completely consistent, and it also makes it more convenient to implement multi-task RL training. The multi-task RL training command is shown below:

```

FPS_MAX_FRAMES=80
QWEN_PATH=Qwen2-VL-7B-zero3-offload-mix_sft_data
HF_DATASET=/rl_data/filter_GQA.json
DS_CONFIG="local_scripts/zero3_offload.json"
CUDA_VISIBLE_DEVICES=0,1,2,3,4,5,6,7 torchrun
--nproc_per_node="8" \
src/open_rl/grpo_clip_tgrpo_qa.py \
--reward_funcs iou accuracy format \
--use_vllm false \
--output_dir ${OUTPUT_DIR} \
--model_name_or_path ${QWEN_PATH} \
--dataset_name ${HF_DATASET} \
--max_prompt_length 16384 \
--max_completion_length 768 \
--per_device_train_batch_size 1 \
--gradient_accumulation_steps 1 \
--learning_rate 2e-6 \
--lr_scheduler_type "cosine" \
--weight_decay 0.01 \
--logging_steps 1 \

```

```

--bf16 true \
--gradient_checkpointing true \
--attn_implementation flash_attention_2 \
--min_pixels 3136 \
--max_pixels 501760 \
--num_train_epochs 1 \
--run_name ${RUN_NAME} \
--save_steps 50 \
--save_total_limit 10 \
--save_only_model false \
--temporal false \
--len_control false \
--use_std true \
--use_length_norm true \
--report_to swanlab \
--beta 0.04 \
--max_grad_norm 5 \
--temperature 1.0 \
--num_generations 8 \
--deepspeed ${DS_CONFIG} \

```

C. Model Inference Details

Inference Complexity. As shown in Fig. 2 in the paper, keyframe localization and temporal grounding can be processed in parallel. We tested on a H20 GPU. The model requires three calls during inference: key element output (~4.6s), frame retrieval (~2.6s), temporal grounding (~4.5s), and the final response (~2.3s). Therefore, the final time is: $\max(4.5, (4.6 + 2.6)) + 2.3 = 9.5\text{s}$. Compared to the non-reasoning model Qwen2.5-VL-Ins, the inference time increased from 2.3s to 9.5s, but it is still significantly lower than the baseline reasoning model Video-R1’s 25s, as shown in Tab. 3.

Latency and computational cost. We added detailed comparative information for different models, such as training and inference latency, FLOPs, and throughput in Tab. 3. Large FLOPs do not mean slow execution; our model is faster than Video-R1 in training and inference.

Table 3. Latency metrics.

Model	Train	Infer	TFLOPs	TP
Qwen2.5-VL	-	2.3s	130	0.43
Video-R1	151s	25.0s	130	0.04
Ours	86s	9.5s	130+13+130	0.11

Offline Inference and Online Inference. Regarding offline inference, it mainly applies to video benchmarks. To facilitate quick and convenient evaluation, for each query and video in the test set, the model can first perform inference on the key intervals and key frames, and then store the positions of these intervals and frames. During the final question-answer inference process, in the stage of pro-

cessing video frames, the key intervals and key frames are sorted and merged. Finally, after the model understands these frames, it outputs the answer. For online inference, the model first performs parallel reasoning to identify the key intervals and key frames. Then, these video frames are input into the model again to obtain the final answer.

D. Qualitative Comparison

We show the inference process comparing the base model and our proposed VideoReasoner in Fig. 1 and Fig. 2. During the reasoning process, the validity of the key intervals and key frames is demonstrated, as shown in Fig 3 and Fig 4. Whether using the Instruct model or the reasoning model, using the default uniform sampling will lead to incorrect analysis and answers. However, our model can obtain the key frames, and this information assists the model in making correct analysis and answers.

References

- [1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 1

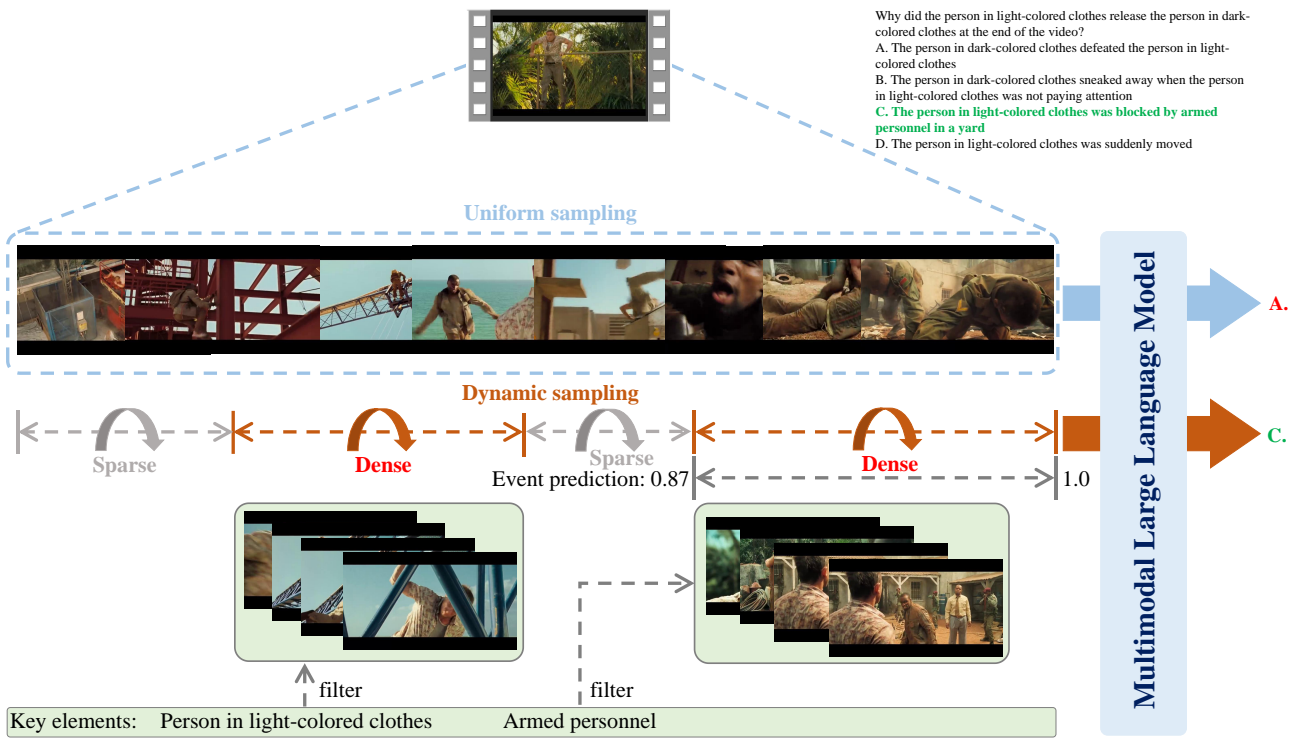


Figure 1. Qualitative comparison between Qwen2-VL-7B-Base and proposed VideoReasoner.

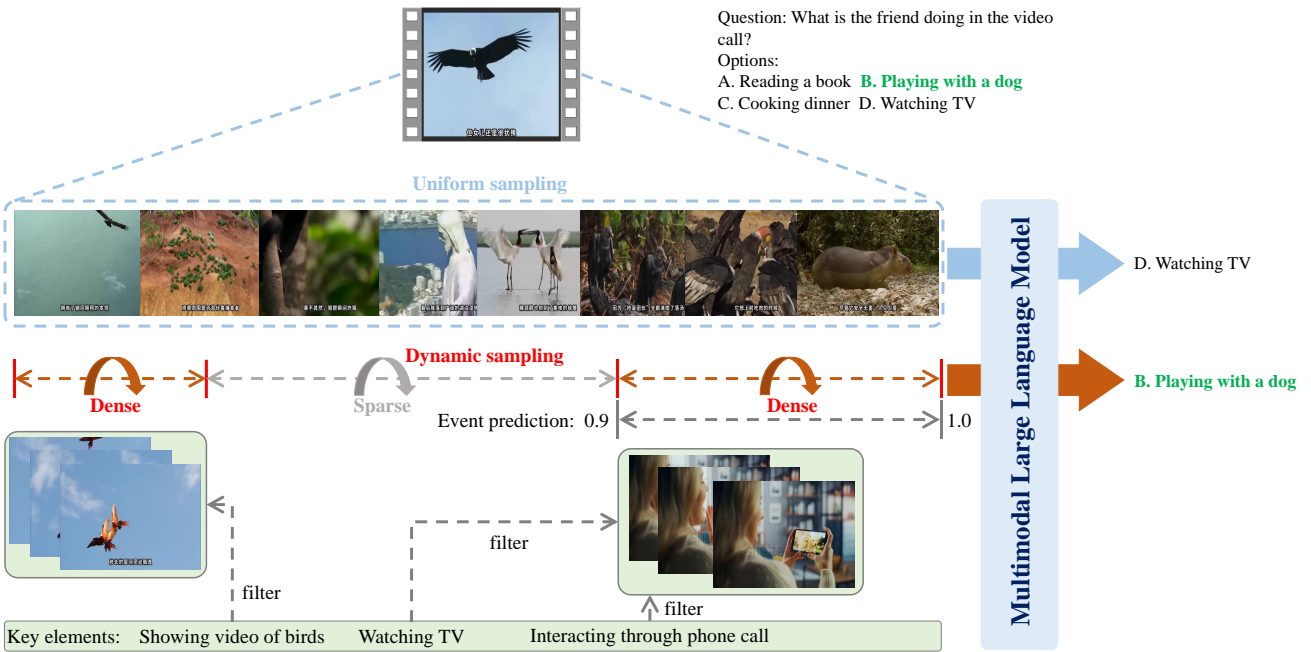
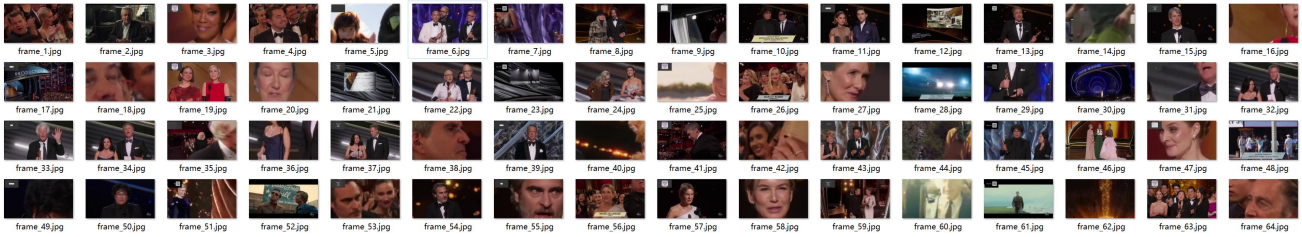


Figure 2. Qualitative comparison between Qwen2-VL-7B-Base and proposed VideoReasoner.



Question: For each award, the number of presenters is different. What are the maximum and minimum numbers of presenters in the video?

Options:

A. 5, 1 B. 2, 1 C. 4, 0 D. 3, 1

Qwen2-VL-7B-Base: **D. 3, 1**

Qwen2-VL-7B-Base+VideoReasoner: **B. 2, 1**



Question: From how other cats treat the cat with the white cane, we can infer about its status?

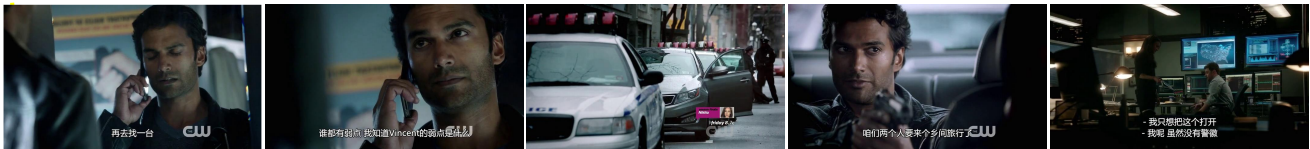
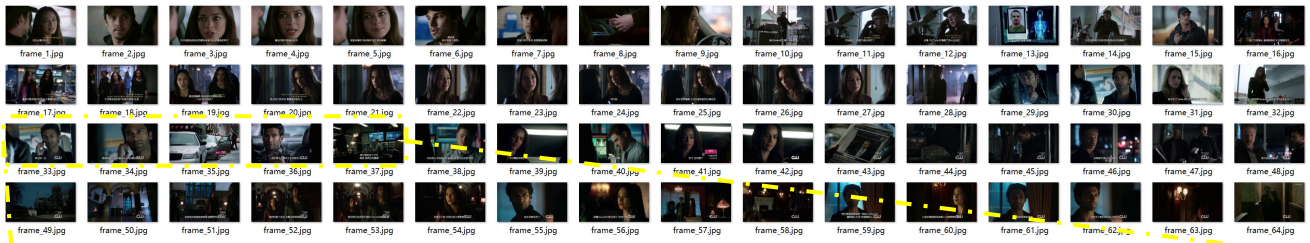
Options:

A. There is no such thing as status for cats B. Its status is very high
 C. Its status is very low D. Its status is similar to theirs

Qwen2-VL-7B-Base: **C. Its status is very low**

Qwen2-VL-7B-Base+VideoReasoner: **B. Its status is very high**

Figure 3. Qualitative comparison between Qwen2-VL-7B-Base and proposed VideoReasoner.



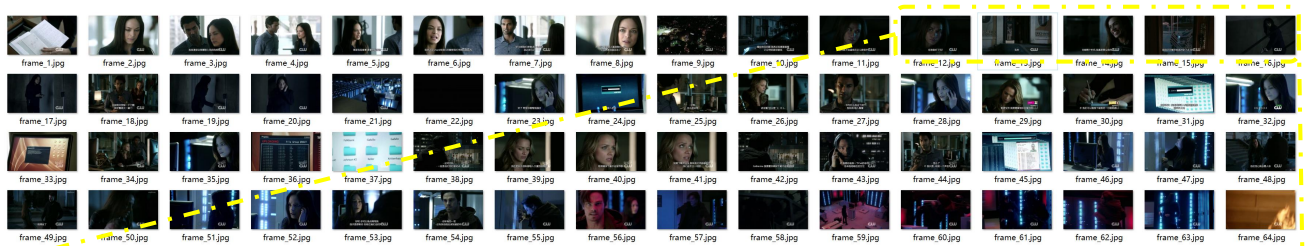
Question: After the woman in the red coat enters the car, what happens to her?

Options:

- A. Car accident
- B. She is hit
- C. She is scolded
- D. She is threatened with a gun

Qwen2.5-VL-7B+Video-R1: <think>Let's break down the sequence of events after the woman in the red coat enters the car:\n1. The woman in the red coat walks towards the car.\n2. She gets into the car.\n3. The scene shifts to show a man in a black jacket standing next to the car.\nNow, let's consider each option:\nA. Car accident - There is no indication of an accident in the video. The woman simply gets into the car and there is no sudden movement or collision.\nB. She is hit - There is no evidence of the woman being hit. The scene does not show any physical altercation or injury.\nC. She is scolded - There is no scolding in the video. The man does not appear to be reprimanding the woman.\nD. She is threatened with a gun - The man in the black jacket is holding a gun, and he appears to be pointing it at the woman. This suggests that she is indeed threatened with a gun.\nGiven the visual cues and the context of the scene, the most accurate description of what happens to the woman in the red coat is that she is threatened with a gun.\nFinal answer: D</think>\n<answer>A. Car accident</answer>

Qwen2.5-VL-7B+Ours: D. She is threatened with a gun



Question: Why does the woman in leather clothing want to attract attention?

Options:

- A. To complete a mission
- B. For fun
- C. To help another woman sneak into the house
- D. Unintentionally

Qwen2.5-VL-7B+Video-R1: <think>The question asks why the woman in leather clothing wants to attract attention. Let's break down the options:\nA. To complete a mission - This seems plausible because the context suggests she is involved in some sort of covert operation or task. Her actions and the environment around her indicate a purposeful intent.\nB. For fun - While it's possible she could be attracted to attention for fun, there's no clear evidence in the video that supports this. The setting and her demeanor suggest a more serious objective.\nC. To help another woman sneak into the house - This is also a possibility. The woman in leather clothing might be using her attention to distract or draw attention away from the other woman, allowing her to enter undetected.\nD. Unintentionally - This is unlikely because her actions and the context suggest a deliberate intention to attract attention.\nConsidering the context and the details in the video, the most logical reason for her to want to attract attention is to complete a mission. The setting and her actions align with a mission-oriented goal rather than personal enjoyment or unintentional distraction.\nAfter analyzing the options, the most accurate and supported reason is A.</think>\n<answer>A. To complete a mission</answer>

Qwen2.5-VL-7B+Ours: C. To help another woman sneak into the house

Figure 4. Qualitative comparison between Video-R1 and our proposed RL training method.